Co uchwała mówi o budżecie?

Mieszkaniec musi odnaleźć właściwą uchwałę, jej zmianę i odpowiedni paragraf w rozproszonych dokumentach.

Jedno pytanie po polsku	RADAR wskazuje zapis w dokumencie, jego datę, stronę i odnośnik do oryginału.
Porównanie wersji	Użytkownik może sprawdzić, jak zmieniła się ta sama pozycja budżetowa w kolejnej uchwale.
55 pozycji źródłowych	Korpus pilotażowy obejmuje 294 fragmenty. Źródła to BIP Krotoszyna i eSesja.
Zakres pilotażu	Niezależny projekt techniczny. Wyniki pochodzą z testów aplikacji. Badanie z mieszkańcami pozostaje do przeprowadzenia.

Jak powstaje odpowiedź

Wyszukiwanie źródeł, wybór fragmentów przez Bielika i kontrola wyniku na serwerze.

1. Miasto i dokument

Serwer sprawdza host, miasto oraz identyfikator dokumentu. Korpus Krotoszyna ma własny zakres danych.

2. Wyszukiwanie

BM25 i reguły zawężają fragmenty według treści pytania, okresu i wersji dokumentu.

3. Bielik Minitron 7B

Polski model GGUF Q4_K_M na NVIDIA T4 wskazuje fragmenty w przekazanym kontekście. Serwer sprawdza identyfikatory i zwraca oryginalne cytaty.

4. Liczby i odmowy

Porównywarka liczy różnice w groszach dla tej samej podstawy. Jednoznaczne pozycje RC i odmowy mogą ominąć model.

Przykład z Krotoszyna

Marcowa zmiana budżetu 2026, uchwała XXXII/248/2026. Poniżej wyniki zweryfikowane na RC.

263 562 056,77 zł	Planowane wydatki ogółem. § 2 w uchwale, s. 1
5 060 421,37 zł	Deficyt budżetu. § 3 w uchwale, s. 1
+6 457 005,89 zł	Zmiana planu wydatków ogółem względem pierwotnej uchwały 2026. Różnicę liczy serwer.
Odmowa przy braku danych	Pytanie o faktycznie wydane środki wymaga zweryfikowanych danych z wykonania budżetu. Uchwały określają plan.

48 z 50 przypadków

Historyczny benchmark techniczny, 13 listopada 2026. Ten sam zestaw uruchomiono w trzech wariantach.

Wariant	Odpowiedzi	Odmowy	Razem
BM25	5/30	0/20	5/50
BM25 z regułami	17/30	20/20	37/50
Pełny RADAR z 7B	**28/30**	**20/20**	**48/50**

Zaliczenie odpowiedzi wymagało oczekiwanego faktu i źródła. Zaliczenie odmowy wymagało braku odpowiedzi przy braku podstawy. Pełny przebieg: 0 błędów HTTP.

Pełny RADAR: 15/17 trafnych świeżych wywołań 7B, 13/13 trafień cache, 20/20 odmów przed inferencją.

Czas odpowiedzi, koszt i walidacja RC

Opóźnienia i koszt pochodzą z historycznego uruchomienia rzeczywistego modelu na NVIDIA T4.

Świeże wywołania · n=17	Mediana 1,674 s, P95 47,330 s, mediana 41,08 tokenu/s. Cache odpowiedzi: mediana 225 ms (n=13).
GPU · 0,50 USD/h	Panel Hugging Face wykazał 0,34 USD za 41 minut. Cennik T4.
Poprawiona wersja RC	Dwie regresje i 9/9 nowych przypadków przeszły. Nowy zestaw obejmował 6 cytatów z uchwał i 3 odmowy reguł.
Stan inferencji w RC	API zarządzania T4 zwróciło 403. Świeża próba 7B na RC pozostaje niewykonana.

Demo, ograniczenia i kolejne kroki

Proponowana kategoria: najlepsze użycie SLM poniżej 10B. Kwalifikacja pilotażu zależy od organizatora.

Wyszukiwanie dokumentów	Chronione Preview Vercel udostępnia wyszukiwanie i porównywanie uchwał po zalogowaniu. Generowanie AI jest obecnie niedostępne.
Historyczne wyniki 7B	Publiczna demonstracja zawiera zapis benchmarku, wideo i odnośniki do źródeł. Dostęp bez logowania.
11 odseparowanych miast	Jedna aplikacja rozpoznaje miasto po domenie. Dołączenie kolejnego korpusu wymaga walidacji źródeł i izolacji.
Dalsza walidacja	Świeża próba 7B na RC, kontrola kosztu i dostępności GPU, uzupełnienie źródeł o wykonanie budżetu, test z mieszkańcami. Regulamin.